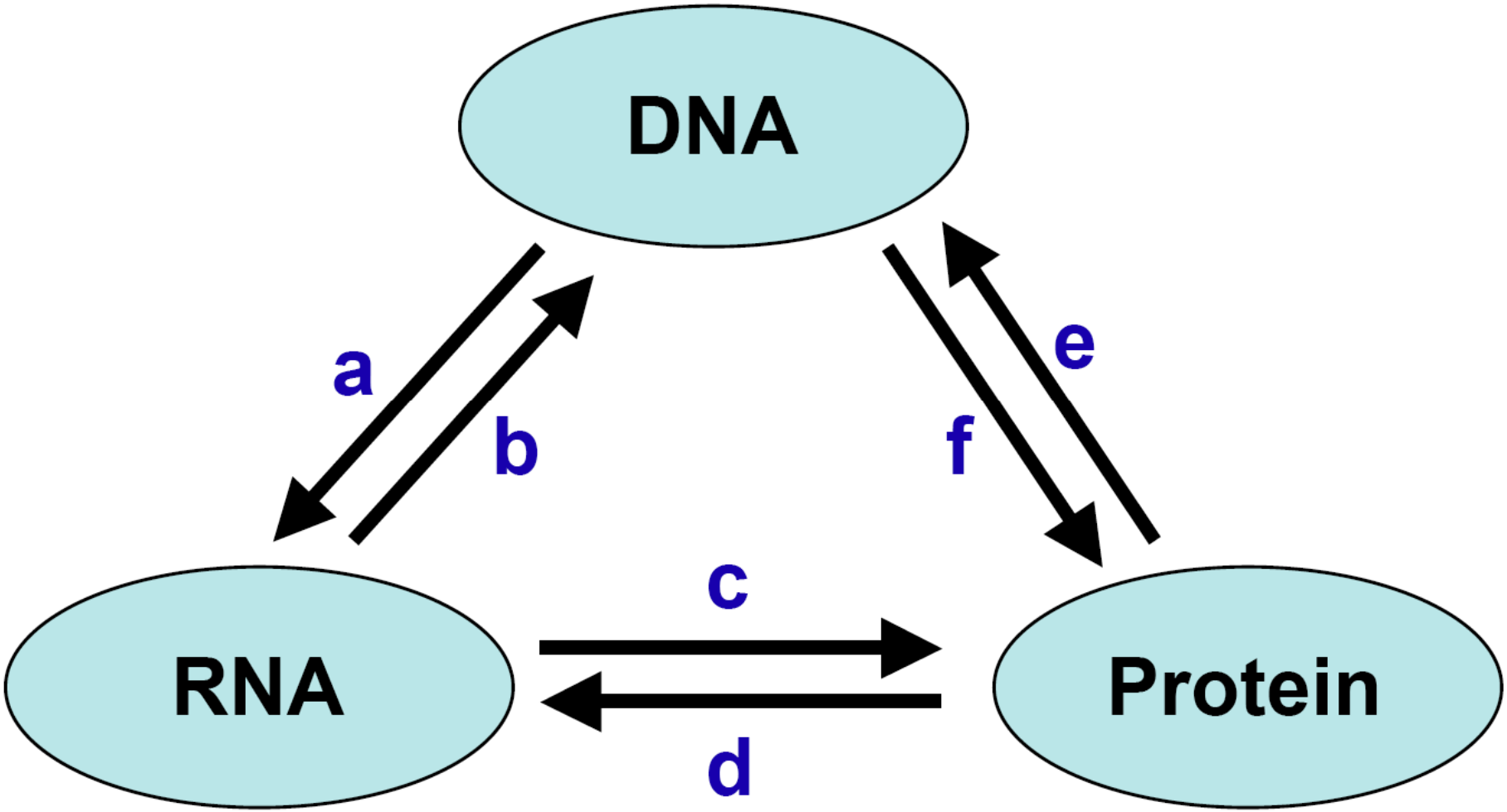
DNA as Biological Information

Rasmus Wernersson
Henrik Nielsen

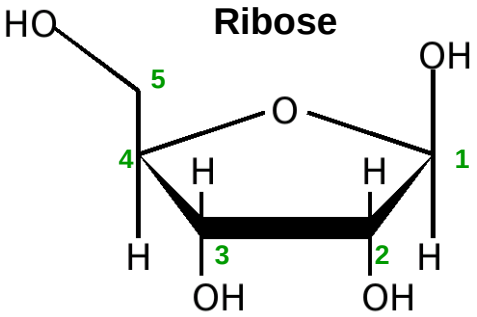
Overview

- Learning objectives
  - About Biological Information
  - A note about DNA sequencing techniques and DNA data
  - File formats used for biological data
  - Introduction to the GenBank database

Information flow in biological systems

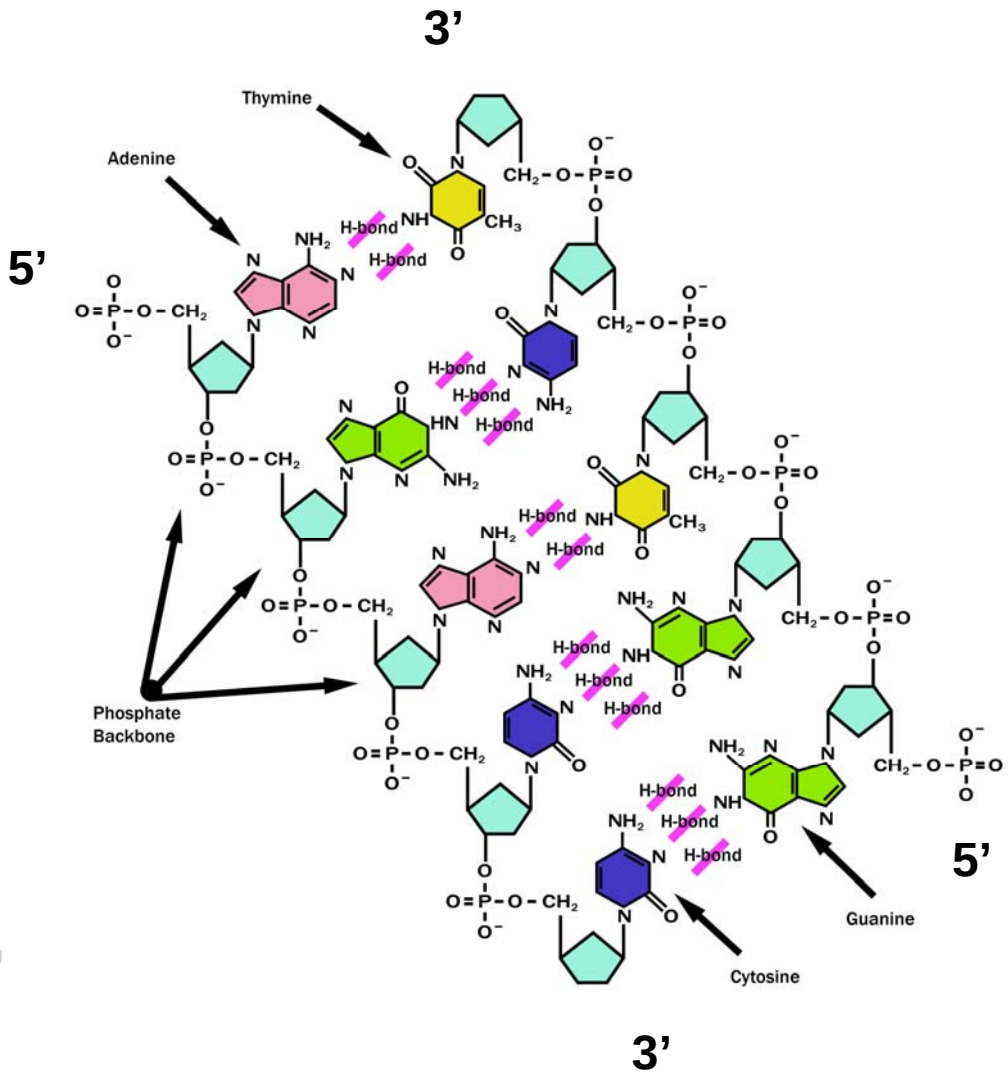


DNA sequences = summary of information



5’ AGCC 3’

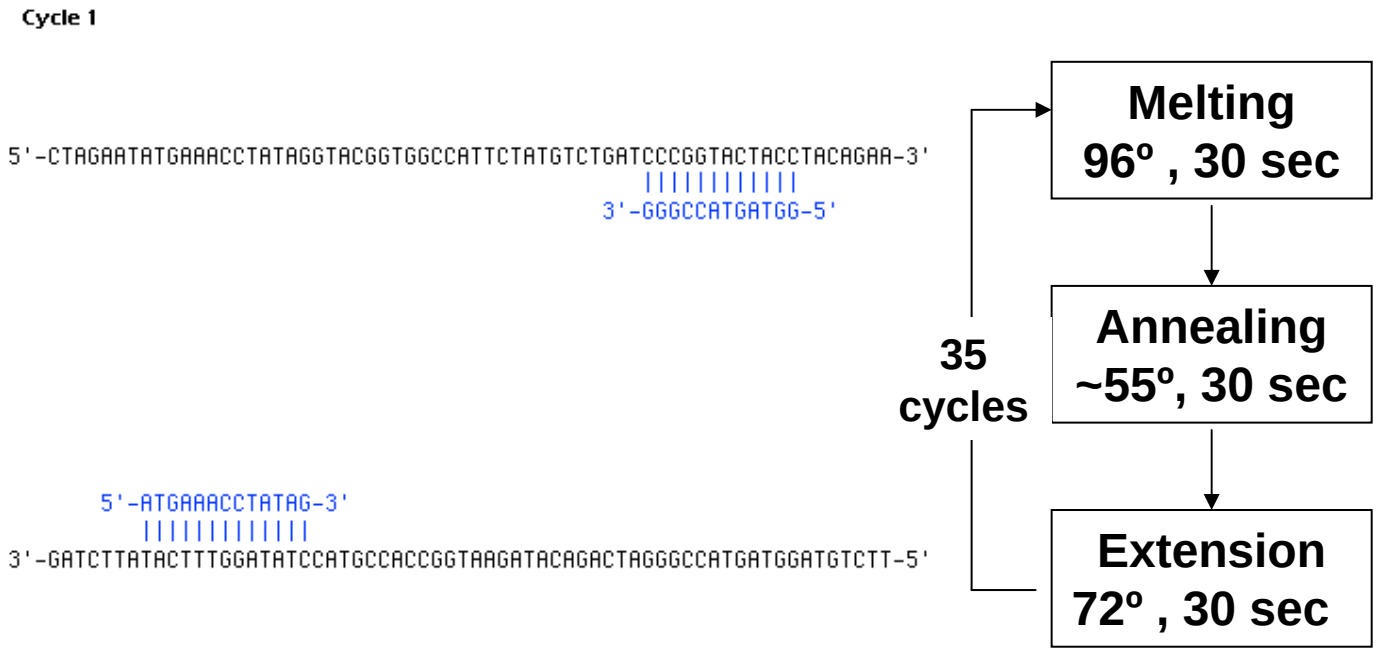
3’ TCGG 5’



5’ ATGGCCAGGTAA 3’

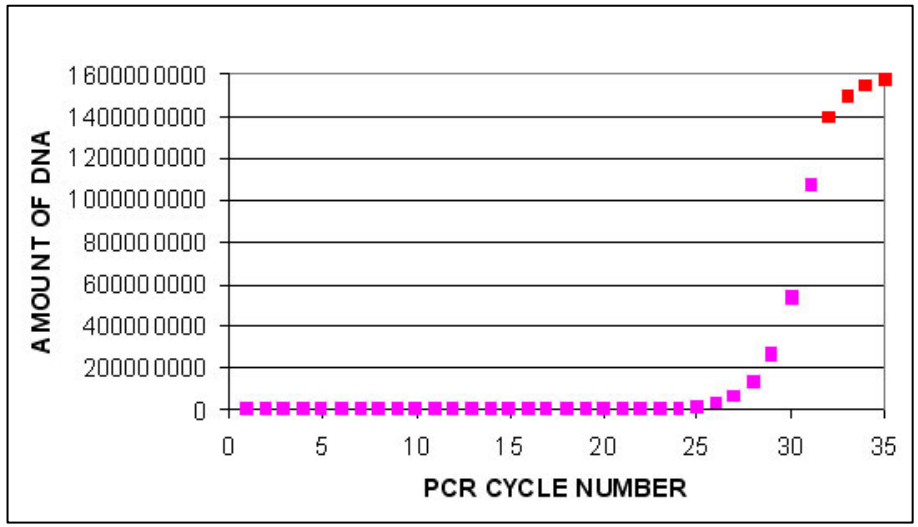
DNA backbone: `http://en.wikipedia.org/wiki/DNA`
(Deoxy)ribose: `http://en.wikipedia.org/`

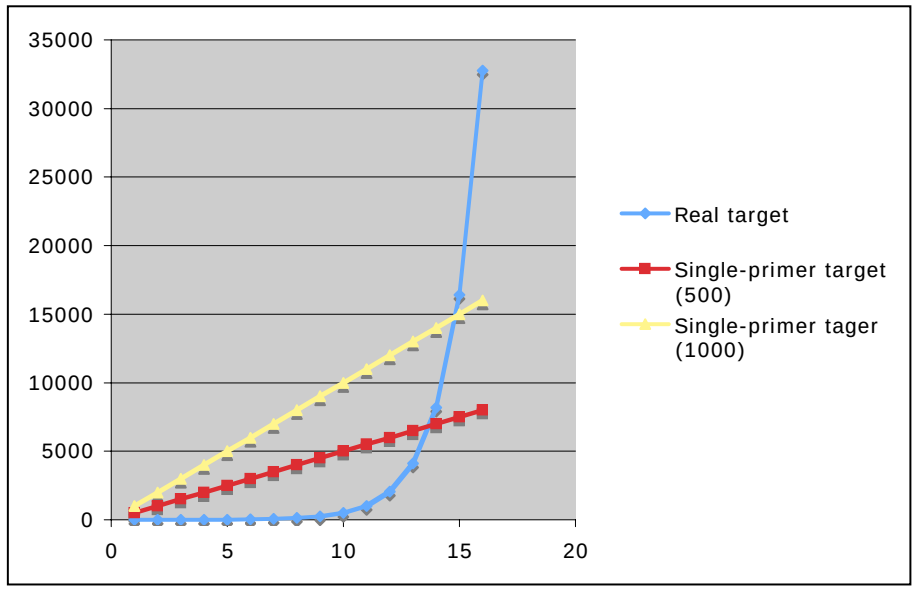
PCR



Animation: http://depts.washington.edu/~genetics/courses/genet371b-aut99/PCR_contents.html

PCR

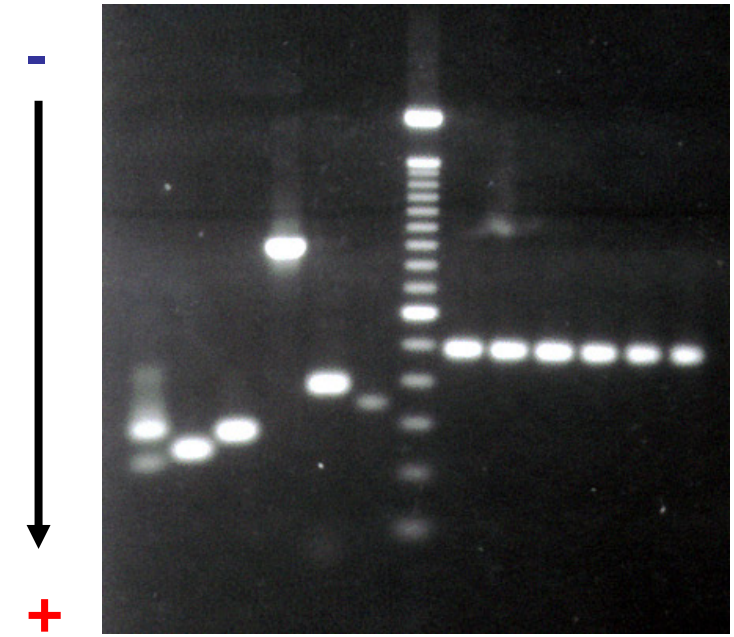


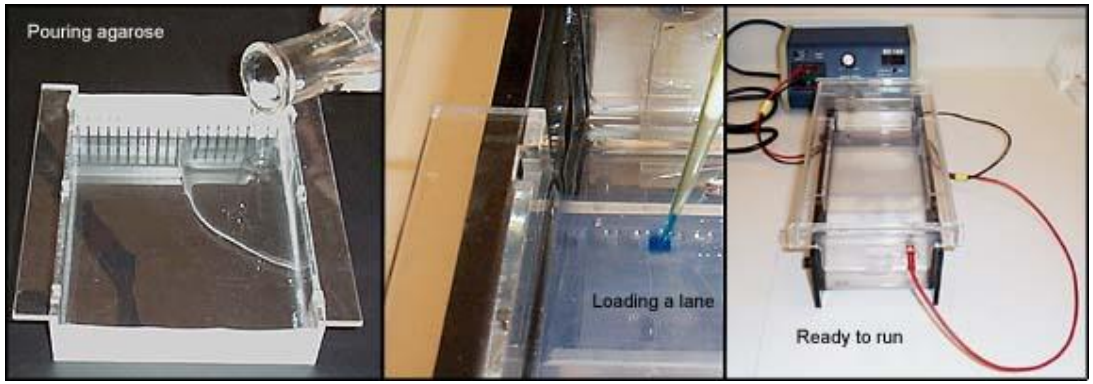


Animation: `http://www.people.virginia.edu/~rjh9u/pcranim.html`
PCR graph: `http://pathmicro.med.sc.edu/pcr/realtime-home.htm`

Gel electrophoresis

- DNA fragments are separated using gel electrophoresis
  - Typically 1% agarose
  - Colored with EtBr or ZybrGreen (glows in UV light).
  - A DNA ”ladder” is used for identification of known DNA lengths.

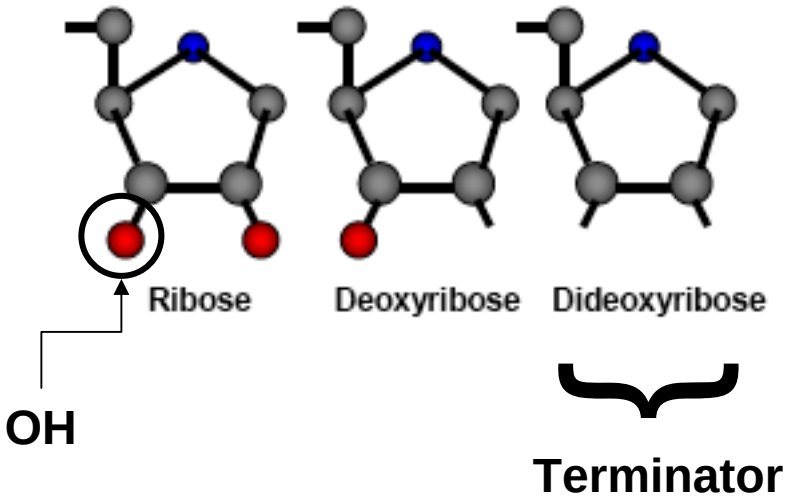


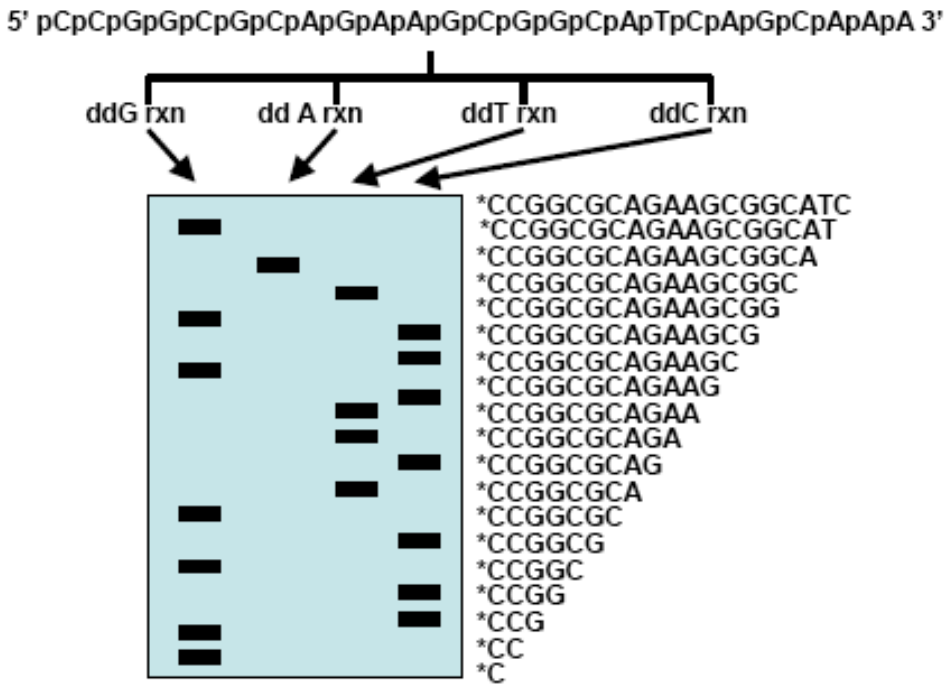


Gel picture: http://www.pharmaceutical-technology.com/projects/roche/images/roche3.jpg

PCR setup: http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/gels/agardna.html

The Sanger method of DNA sequencing



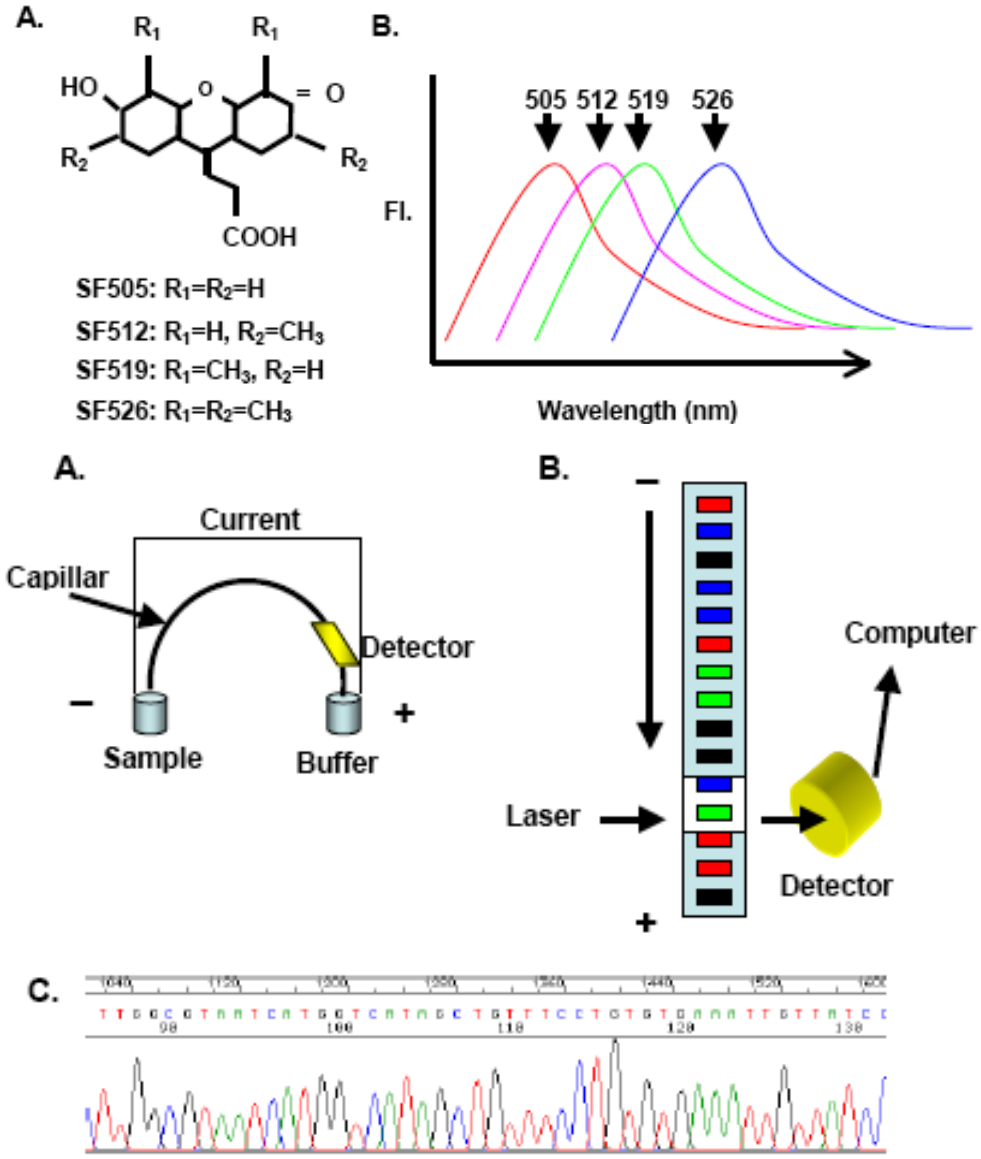


X-ray sequenceing gel

Images: http://www.idtdna.com/support/technical/TechnicalBulletinPDF/DNA_Sequencing.pdf

Automated sequencing

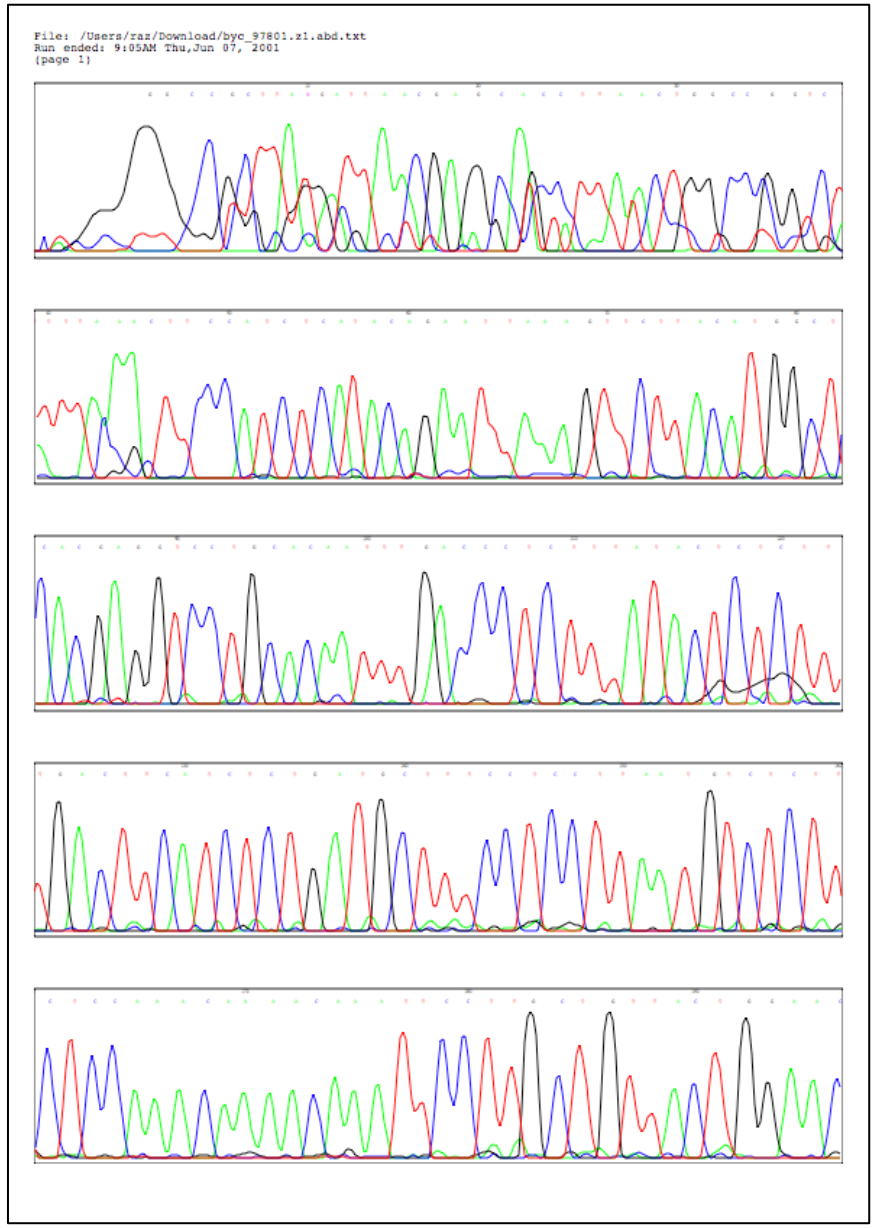
- The major break-through of sequencing has happened through *automation.*
- Fluorescent dyes.
- Laser based scanning.
- Capillary electrophoresis
- Computer based base-calling and assembly.



Images: `http://www.idtdna.com/support/technical/TechnicalBulletinPDF/DNA_Sequencing.pdf`

Handout exercise: ”base-calling”

- Handout: Chromotogram
- Groups of 2-3.
- Tasks:
  - Identify “difficult” regions
  - Identify “difficult” sequence stretches.
  - Try to estimate the best interval to use.

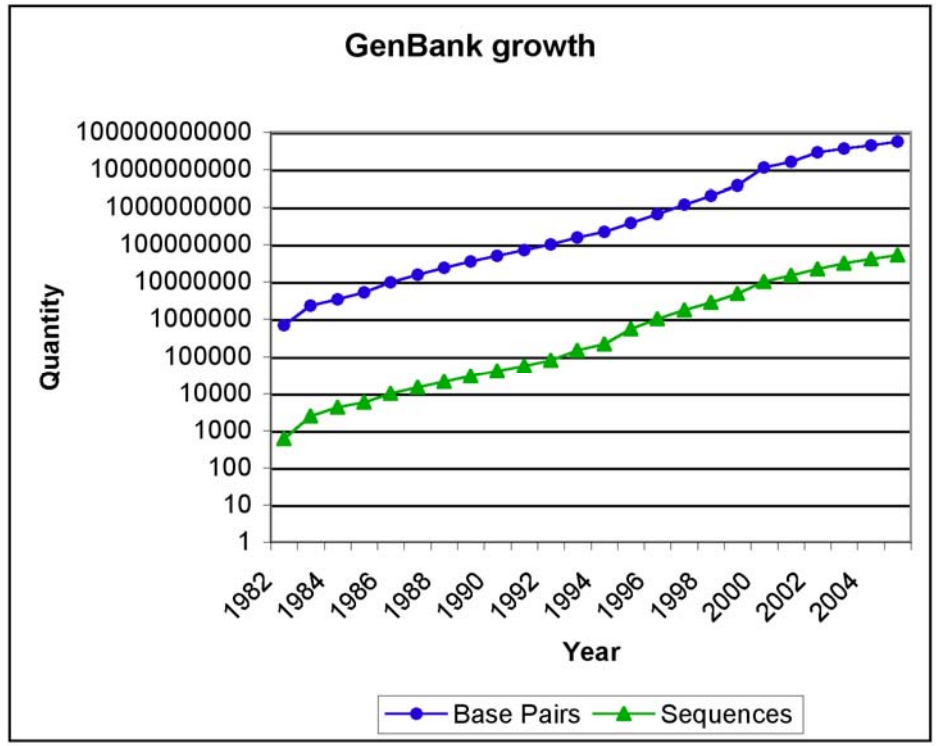


Biological data on computers

- The GenBank database
- File formats
  - FASTA
  - GenBank

NCBI GenBank

- GenBank is one of the main international DNA databases.
- GenBank is hosted by NCBI: ***N**ational **C**enter for **B**iotechnology **I**nformation*.
- GenBank has existed since 1982.
- The database is public - no restrictions on the use of the data within.

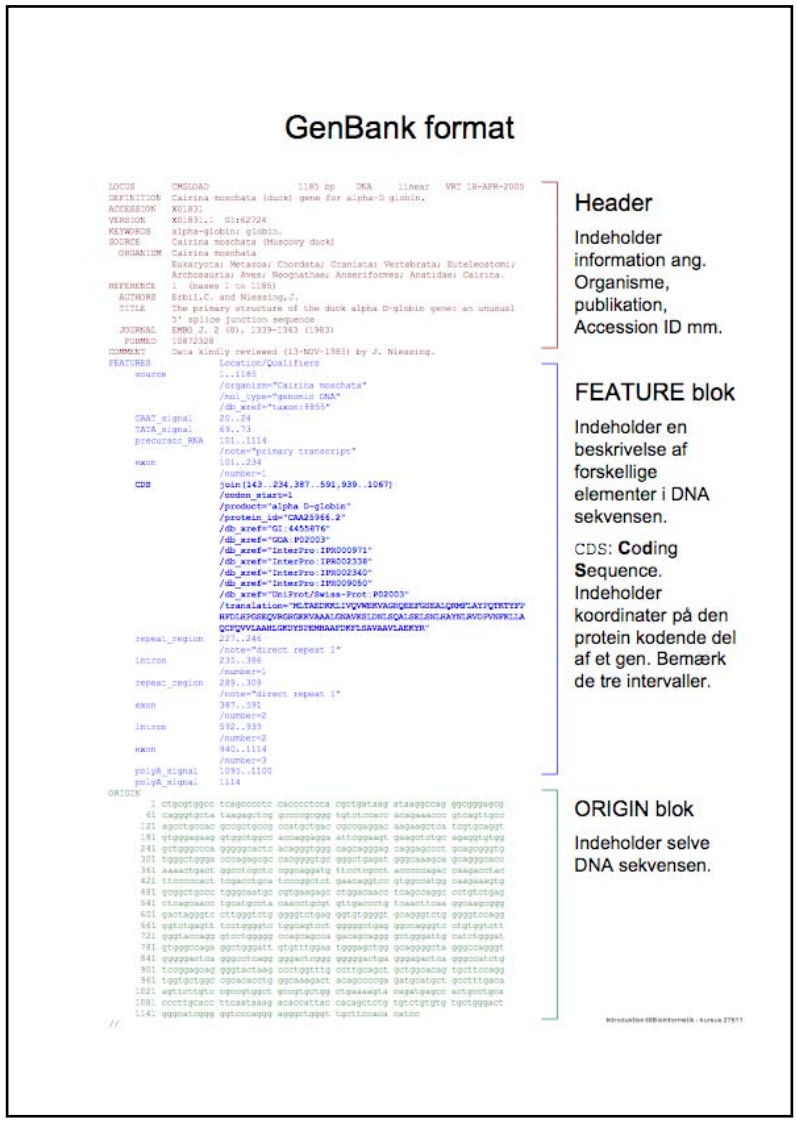


FASTA format

>alpha-D
ATGCTGACCGACTCTGACAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCAC
CCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCA
AGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTTGTTCGGCTGGGGGTCAG
CACTGACCATCCCGCTCCCGCAGCTGTTCACCACCTACCCCCAGACCAAGACCTACTTCC
CCCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGG
CCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCA
GCGACCTGCATGCCTACAACCTGCGTGTCGACCCTGTCAACTTCAAGGCAGGCGGGGGAC
GGGGGTCAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCC
GGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTG
GCCACACACCTGGGCAACGACTACACCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTG
TCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAA
>alpha-A
ATGGTGCTGTCTGCCAACGACAAGAGCAACGTGAAGGCCGTCTTCGGCAAAATCGGCGGC
CAGGCCGGTGACTTGGGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCC
ATCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCCATACTCTCCCCATCCATAACTG
TCCCTGTTCTATGTGGCCCTGGCTCTGTCTCATCTGTCCCCAACTGTCCCTGATTGCCTC
TGTCCCCCAGGTTGTTCATCACCTACCCCCAGACCAAGACCTACTTCCCCCACTTCGACC
TGTCACATGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTGGCGGAGGCACTGGTTG
AGGCTGCCAACCACATCGATGACATCGCTGGTGCCCTCTCCAAGCTGAGCGACCTCCACG
CCCAAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCA
GTCTGGCTCCCCTCCTGCACACACCTCTGGCTACCCCCTCACCTCACCCCCTTGCTCACC
ATCTCCTTTTGCCTTTCAGCTGCTGGGTCACTGCTTCCTGGTGGTCGTGGCCGTCCACTT
CCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGTGTGTGCCGTGGG
CACCGTCCTTACTGCCAAGTACCGTTAA

(Handout)

GenBank format



(handout)

- Originates from the GenBank database.
- Contains both a DNA sequence and annotation of feature (e.g. Location of genes).

GenBank format - HEADER

```
LOCUS       CMGLOAD                 1185 bp    DNA     linear   VRT 18-APR-2005
DEFINITION  Cairina moschata (duck) gene for alpha-D globin.
ACCESSION   X01831
VERSION     X01831.1  GI:62724
KEYWORDS    alpha-globin; globin.
SOURCE      Cairina moschata (Muscovy duck)
  ORGANISM  Cairina moschata
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE   1  (bases 1 to 1185)
  AUTHORS   Erbil,C. and Niessing,J.
  TITLE     The primary structure of the duck alpha D-globin gene: an unusual
            5' splice junction sequence
  JOURNAL   EMBO J. 2 (8), 1339-1343 (1983)
   PUBMED   10872328
COMMENT     Data kindly reviewed (13-NOV-1985) by J. Niessing.
```

GenBank format - ORIGIN section

```
ORIGIN
        1 ctgcgtggcc tcagcccctc cacccctcca cgctgataag ataaggccag ggcgggagcg
       61 cagggtgcta taagagctcg gccccgcggg tgtctccacc acagaaaccc gtcagttgcc
      121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
      181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
      241 gctgggccca gggggcactc acagggtggg cagcagggag caggagccct gcagcgggtg
      301 tgggctggga cccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
      361 aaaactgact ggcctcgctc cggcaggatg ttcctcgcct acccccagac caagacctac
      421 ttccccccact tcgacctgca tcccggctct gaacaggtcc gtggccatgg caagaaagtg
      481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
      541 ctcagcaacc tgcatgccta caacctgcgt gttgaccctg tcaacttcaa ggcaagcggg
      601 gactagggtc cttgggtctg ggggtctgag ggtgtggggt gcagggtctg ggggtccagg
      661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
      721 gggtaccagg gtcctggggg ccagcagcca gacagcaggg gctgggattg catctgggat
      781 gtgggccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
      841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
      901 tccggagcag gggtactaag ccctggtttg ccttgcagct gctggcacag tgcttccagg
      961 tggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
     1021 agttcttgtc cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
     1081 cccttgcacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
     1141 gggcatcggg ggtcccaggg agggctgggt tgcttccaca catcc
//
```

GenBank format - FEATURE section

```
FEATURES             Location/Qualifiers
     source          1..1185
                     /organism="Cairina moschata"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:8855"
     CAAT_signal     20..24
     TATA_signal     69..73
     precursor_RNA   101..1114
                     /note="primary transcript"
     exon            101..234
                     /number=1
     CDS             join(143..234,387..591,939..1067)
                     /codon_start=1
                     /product="alpha D-globin"
                     /protein_id="CAA25966.2"
                     /db_xref="GI:4455876"
                     /db_xref="GOA:P02003"
                     /db_xref="InterPro:IPR000971"
                     /db_xref="InterPro:IPR002338"
                     /db_xref="InterPro:IPR002340"
                     /db_xref="InterPro:IPR009050"
                     /db_xref="UniProt/Swiss-Prot:P02003"
                     /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
                     HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLA
                     QCFQVVLAAHLGKDYSPEMHAAFDKFLSAVAAVLAEKYR"
     repeat_region   227..246
                     /note="direct repeat 1"
     intron          235..386
                     /number=1
     repeat_region   289..309
                     /note="direct repeat 1"
     exon            387..591
                     /number=2
     intron          592..939
                     /number=2
     exon            940..1114
                     /number=3
     polyA_signal    1095..1100
     polyA_signal    1114
```

Exercise: GenBank

- Work in groups of 2-3 people.
- The exercise guide is linked from the course programme.
- Read the guide carefully - it contains a lot of information about GenBank.